



US Data Initiatives

Beth Plale

Indiana University

Rebecca Koskela

University of New Mexico

US Data Initiatives

- Any summary of US activity is by nature a sampling
- Other good activities:
 - Earth Science Grid (ESG)
 - NSF Big Data awardees (summer 2012)
 - NIH activities
 - Domain activities
 - E.g., Int'l Virtual Observatory Alliance
 - Well known standards activities
 - E.g., Open Geospatial Consortium (OGC)



TERRA POPULUS

Integrated Data on Population and Environment

Steven Ruggles
University of Minnesota



Primary Objective

Make data with different formats from different scientific domains easily interoperable

- Population microdata
- Government land-use statistics
- Land cover data from satellite imagery
- Historical climate records (temperature, precipitation, cloud cover)

Relationship



H910000240000000088001001000220100
P910000020101032120010010010011504
P910000010201036220010010010011999
P910201000301011220060010010011999
P910201000301009120060010010011999
P910201000301007120060010010011999
P910201000301006120060010010011999
P910201000301004220060010010011999



Facebook has data on
800 million people

Forbes

TECH | 12/07/2011 @ 12:46PM | 1,180 views

Future of Advertising is Data: Facebook's Numbers Staggering - 800 million users; 500 million active per day

+ Comment now

Facebook's insanely rapid growth is well documented as well as their initial public offering (and insane valuation). It was reported last month that Facebook is looking at the April to June 2012 timeframe to raise \$10 billion in an IPO. As Facebook ramps up for their IPO they are riding high with an impressive user base. To me the most impressive milestone around Facebook is the impressive stats around their user base.

facebook

Inside the Facebook Numbers

Total Registered Users: 800 million

Total Active Daily Users: 500 million

Total Mobile Users: 350 million

Apps & Websites Integrated in Facebook Platform: 7 million

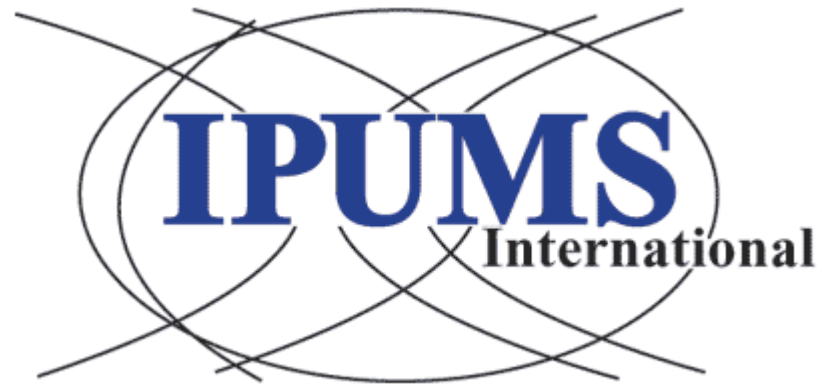
Total Photo's Shared on Halloween: > 1 billion

Facebook's scale and user number are quite impressive. Can they keep the growth up? Just like Google the more Internet users that are online the more growth Facebook can achieve.

Facebook's IPO Valuation Insane – Maybe Not?

Facebook's valuation estimated at \$100 billion is insane by any standard, but

We have data on
912 million people



USA	165
International	481
Historical	266
Total	912



Preservation

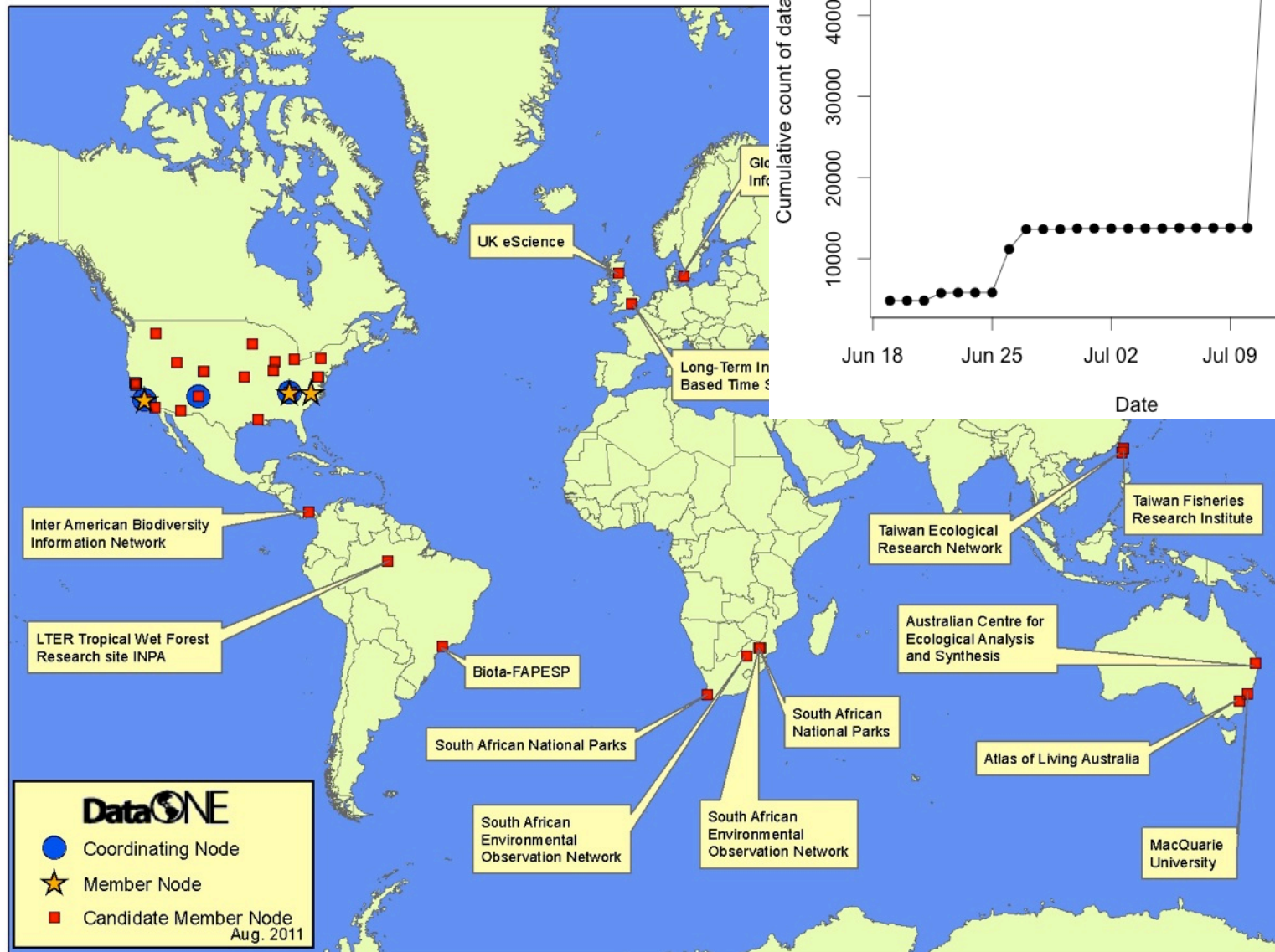


DataONE: Supporting Scientific Data Preservation, Discovery, and Innovation

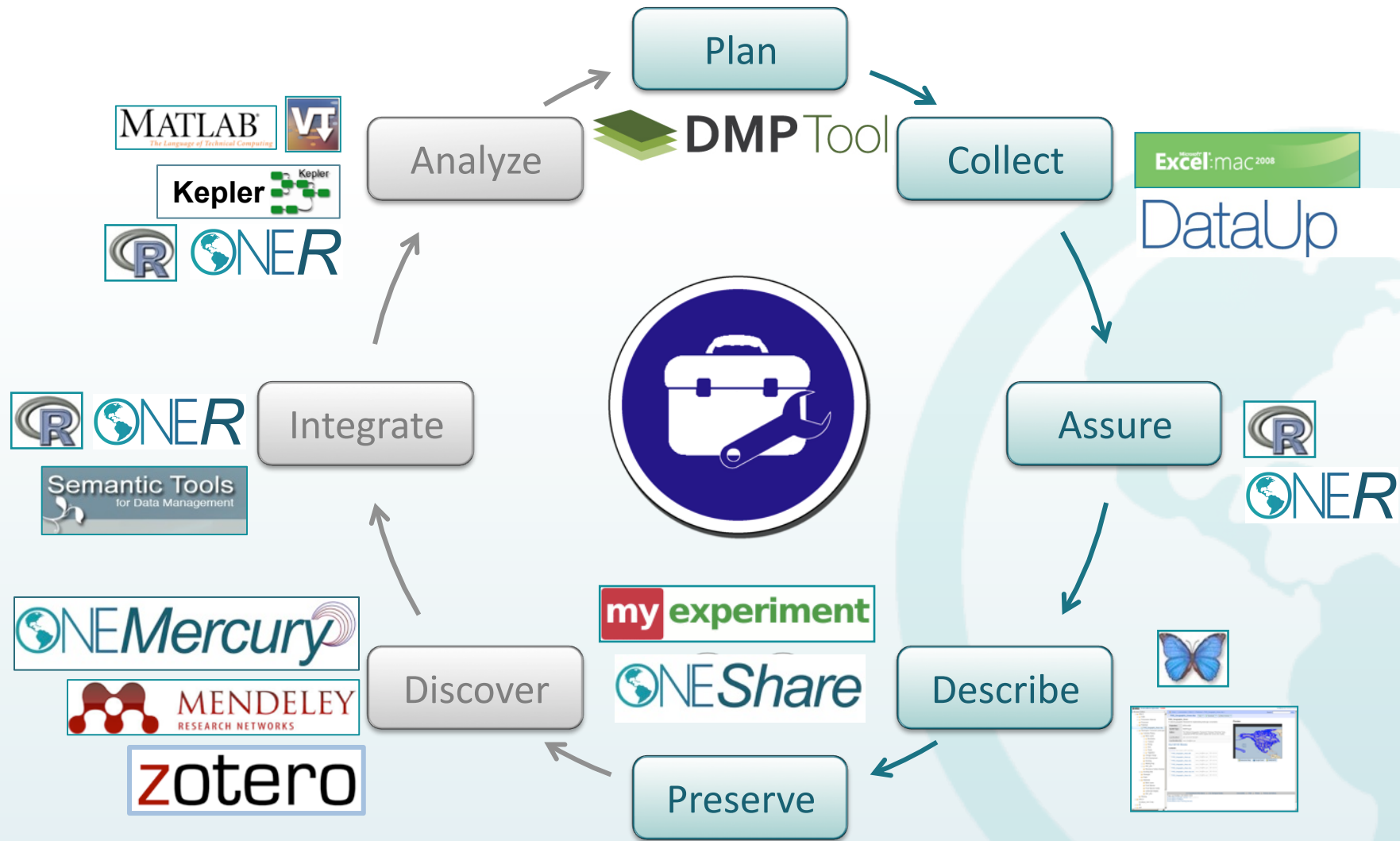
Bill Michener
University of New Mexico



The DataONE Federation



Investigator Toolkit Support



Federated Resources

Member Nodes (MNs)

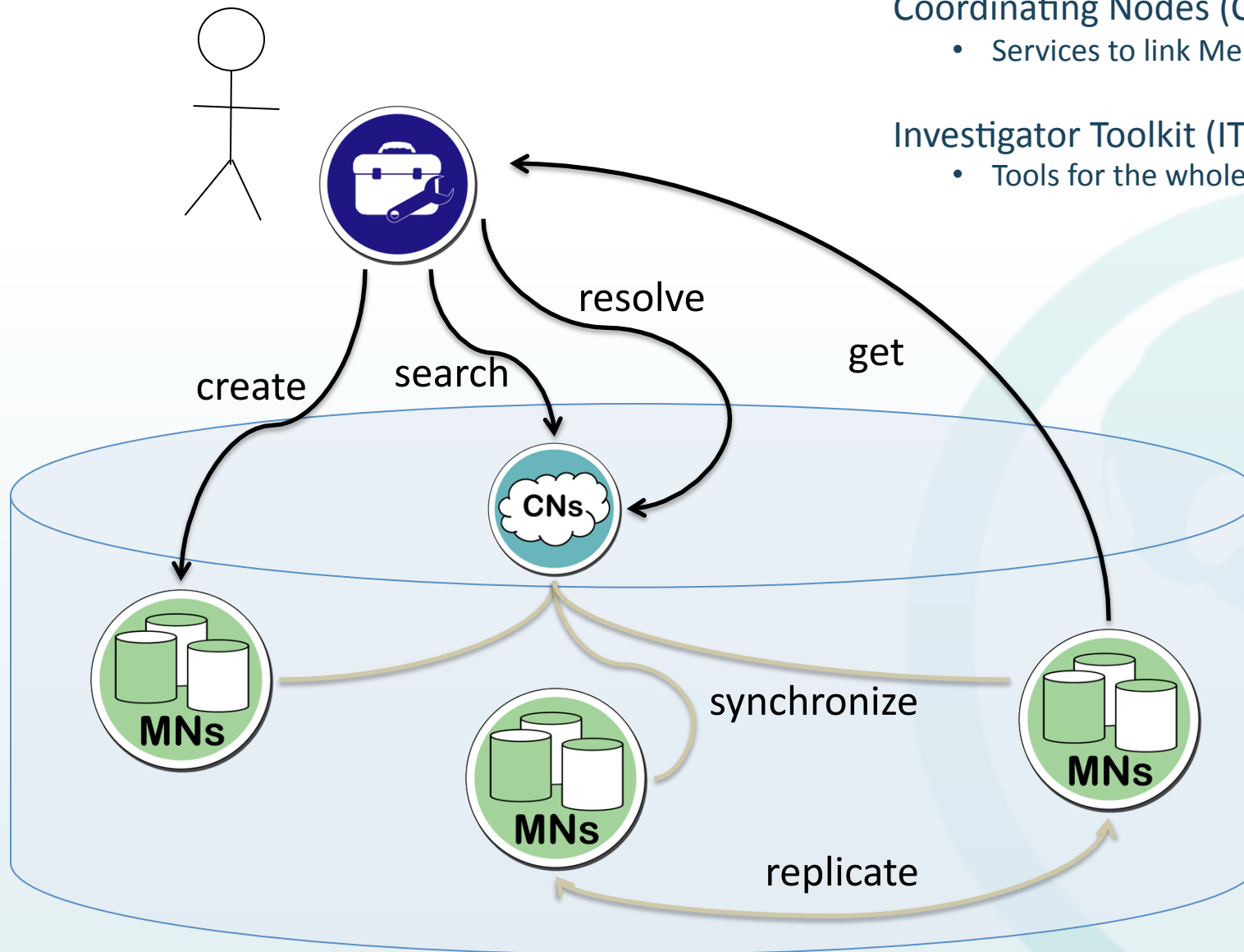
- Heart of the federation
- Harness the power of local curation

Coordinating Nodes (CNs)

- Services to link Member Nodes

Investigator Toolkit (ITK)

- Tools for the whole data lifecycle





Data Conservancy

Sayed Chouhardy, Johns Hopkins

- Data Conservancy (DC) is a community that develops solutions for data preservation and sharing to promote cross-disciplinary re-use.
- DC Service Instance: data centric hardware, software, components, and APIs within an organizational context – installed at Johns Hopkins University and (soon) National Snow and Ice Data Center



DC Service software architecture

DCS HTTP API Endpoints

Entity
Retrieval API

Data Model
Query API

GQM
Query API

DataStream
API

Lineage
API

Deposit
API

DCS Java APIs

ArchiveStore API

Index API

Query API

Ingest
Framework

Feature
Extraction
Framework

Lineage API

Common Services

ID
Service

ArchiveStore Impl

Fedora

ELM

Index and Query Impl

Solr

GQM

Ingest Impl

FEF Impl

Lineage Impl



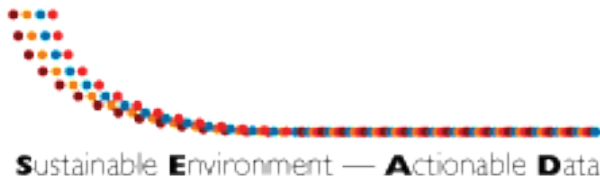
Data Exchange Attributes

- Feature Extraction Framework that atomizes data into constituent parts for indexing, metadata extraction, etc.
- Discipline agnostic data model (inspired by PLANETS project)
- Provenance and Lineage service
- Spatial, temporal and (soon) taxonomic query capabilities
- Sustainability through diverse funding from Johns Hopkins University, direct charges to NSF grants, other grants and community development

SEAD: Sustainable Environment - Actionable Data

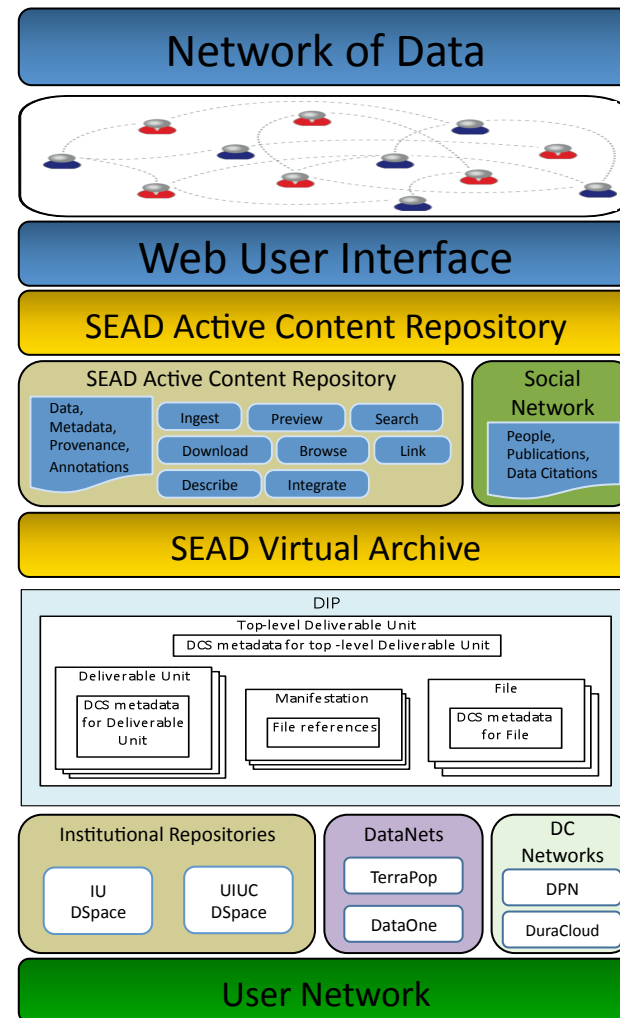
Margaret Hedstrom, Univ Michigan

- Long tail: secure and long term storage for active data along with basic tools and services (discovery, geo-referencing, visualization).
- Active and Social Curation
 - Capture as much metadata as possible automatically to make data actionable.
 - Allow data producers and users to improve the quality, relevance, usability, and usefulness of data through rating, annotation, metadata contribution, and quality control, etc.
- Deploy social networking tools (VIVO) to hasten discovery of people, expertise, data, and publications.



SEAD Stack

- Data-to-user pipeline for data to flow from data producers through *active curation* to one or more *permanent repositories*.
- Bulk ingest of NCED data (1.5 TB; 400,000+ files)
- Profiles to SEAD VIVO for all NCED researchers



Social/Active Interfaces



Praveen Kumar

Professor, Civil and Environmental Engineering, University of Illinois, Urbana, Illinois 61801

[Hydrology](#)

Verified email at illinois.edu

[Homepage](#)

SEAD Home • Data • Collections • Tags • Map • Upload • Administration James Myers (Logout) Search

LiDAR_Raster.jpg

Image Zoom

Download Delete Rerun Extraction Embed

Info

Contributor: Md Aktaruzzaman
 Creator: Praveen Kumar - <http://vivo-vis-test.sls.indiana.edu/>
 Filename: LiDAR_Raster.jpg
 Size: 132.04 KB
 Category: Image
 MIME Type: image/jpeg
 Uploaded: 2012-03-07 10:18
 Image Size: 549x732

License

All Rights Reserved
[Edit](#)

Social

Viewed by 4 people
 Downloaded by 0 people
 1 likes and 0 dislikes
[Unlike](#) [Dislike](#)

Tags

[raster](#) [Remove](#)
[spatial](#) [Remove](#)
[lidar_dem](#) [Remove](#)

[Add](#) [Cancel](#)

Collections

User Specified Information

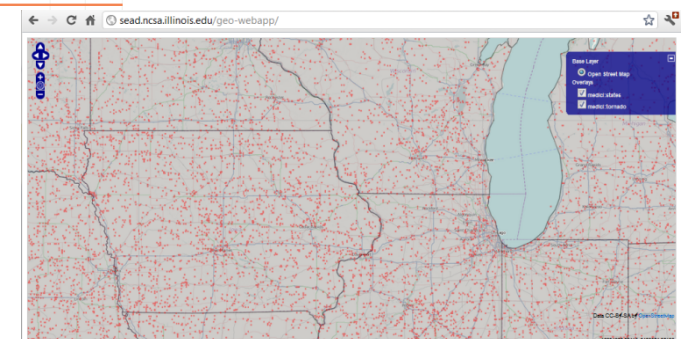
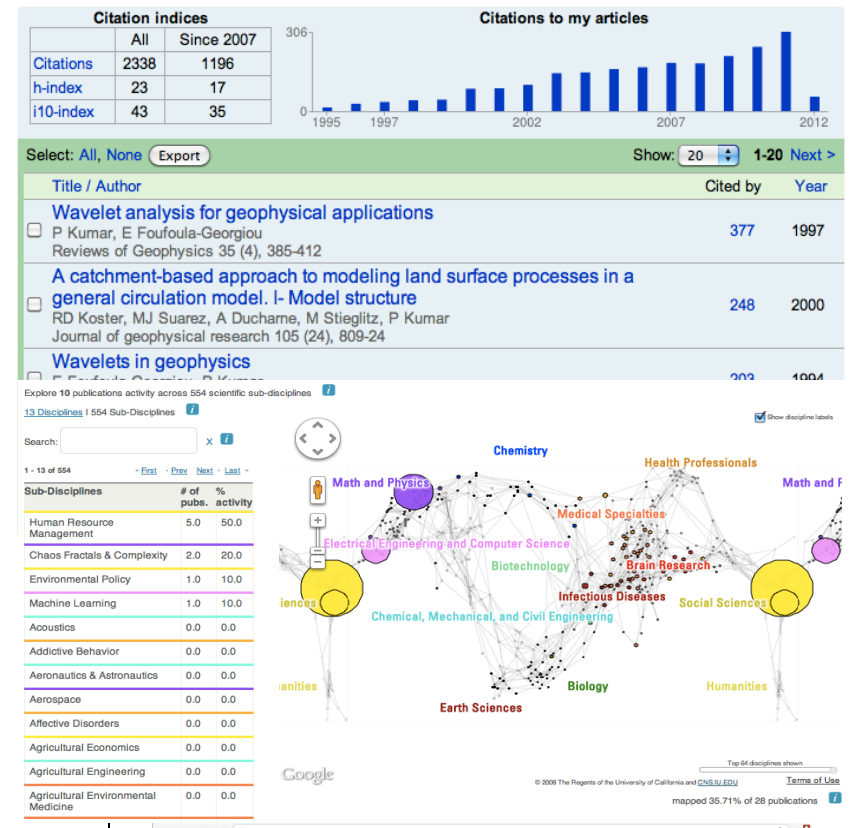
Field	Value	Applies To	Action
Creator	Praveen Kumar - http://vivo-vis-test.sls.indiana.edu/vivo/display/a1020	Document	Edit Remove

Creator [Add](#) [Clear](#)

Extracted Information

Drop files to upload

IMG_1778.JPG:	Uploading...
IMG_1779.JPG:	Uploading...
IMG_1780.JPG:	Uploading...
IMG_1781.JPG:	Uploading...
IMG_1782.JPG:	Uploading...
IMG_1783.JPG:	Uploading...
IMG_1784.JPG:	Uploading...



DataNet Federation Consortium

Data Driven Science

- Implement national data infrastructure
 - Federate existing discipline-specific data management systems to enable national research collaborations
- Enable collaborative research on shared data collections
 - Manage collection life cycle as the user community broadens
- Integrate “live” research data into education initiatives
 - Enable student research participation through control policies

Project

Shared Collection

Processing Pipeline

Digital Library

Reference Collection

Federation

Collection Life Cycle

Reagan W. Moore,
University of North Carolina Chapel
Hill



Policy-based
data management

Build National Infrastructure Through Federation

- Ocean Observatories Initiative, National Climatic Data Center
 - Data grid for oceanography, sensor control, real-time data streams, archive
- CUAHSI, UNC Institute for the Environment, National Climatic Data Center
 - Data grid for hydrology, watershed modeling workflow integration
- CIBER-U (Engineering design, undergraduate education)
 - Digital Library, OOI sensor documents
- Years 3-5
- the iPlant Collaborative
 - Data grid for plant biology, federation with existing biology resources
- Odum Social Science Research Institute
 - DataVerse federation, data archive
- Temporal Dynamics of Learning Center
 - Data grid for cognitive science

Enabling Tools

- Data grid
 - Build shared name spaces for users, files, resources, metadata, rules, procedures
- Soft links
 - Register data from external data management system, accessed through its protocol
- Federated data grids
 - Cross-register users between data management systems
- Workflow integration
 - Register workflows into data grid for storage side procedures
 - Integrate data management workflows with external workflows

Tao of ESIP

- ***We are...***

- **→ Community-driven**

- Members are the authority
- Voluntary
 - No requirements
 - No remuneration
 - “For the good of the order”
- Distributed
 - Geographically
 - Topically
 - Functionally
- Open
 - Collegial
 - Neutral forum

- ***We value...***

- Participation
 - Share your expertise
 - Leverage others’ expertise
 - Encourage free flow of ideas
 - Exposure → opportunities
- Collaboration
 - “Communities of practice”
- Innovation
 - No institutional barriers
 - Results for \$5K!
- ***Hybrid virtual and ‘real’ organization***



Data Related Activities & Accomplishments

- ***Adopted data management principles & practices for:***

- Data producers
- Data intermediaries
- Data users

- ***Investigated identifiers for data***

- Published paper w results
- Testbed activity validated results

- ***Working on provenance & context content standard***

- NASA adopted a version for its missions
- Hope to eventually create an IEEE or ISO standard from it

- ***Data management training***

- In process of publishing ~50 short presentations for scientists

- ***Data quality working group***

- Just starting up again

- ***Discovery cluster work***

- Protocols so simple why wouldn't you use them!
- Geospatial/temporal OpenSearch
- Data and services advertising feeds

- ***Earth Science Collaboratory***

- Goal is to provide a rich data analysis environment
- Access to Earth Science data, tools and services
- Support for collaboration and sharing

- ***Documentation working group***

- Working to understand the different metadata dialects/conventions across ESIP members

Cyberinfrastructure for Social Sciences

- A Workshop organized by University of Washington, UC Berkeley, Internet2 and Microsoft Research
 - Oct 15-16th Seattle <http://www.i2azure.com>

Goals

- Understand the big data challenges facing the social sciences research community
 - Data curation, analytics, collaboration, provenance
- How do we build community supported, financially sustainable data collections and analytics services?
 - Can we build a research data marketplace?
- What is the role of commercial cloud providers?



MISSION

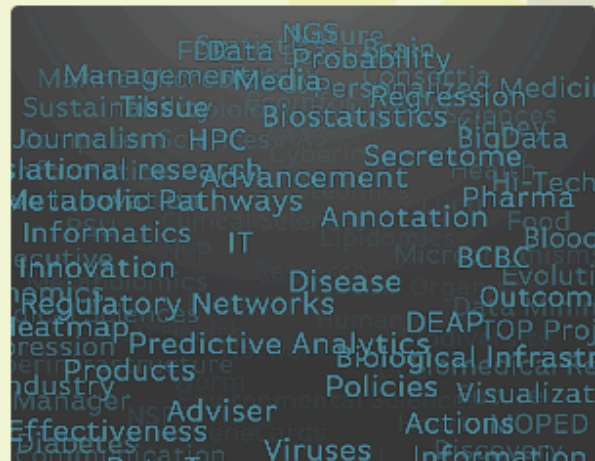
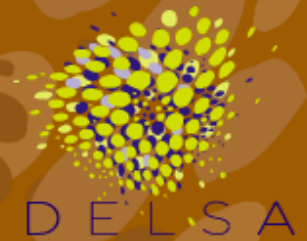
VISION

STRATEGY

Accelerate the impact of

Data-Enabled Life Sciences Research

on the pressing needs of the global society.



UPCOMING EVENTS

DELSA Workshop III

October 9, 2012

Part of the 8th IEEE International
Conference on eScience, Hyatt
Regency Chicago, Chicago, IL

[READ MORE...](#)

RELEVANT READING

Summer 2012 Newsletter

We have many exciting things happening at DELSA!
In this newsletter you will find:

- *NEW NSF funded project (see DELSA Endorsed Projects Update)
- *Upcoming meeting on October 9th (see Events)
- *Highlighted publications by DELSA members (See DELSA in the News)

DELSA Strategy

- Identifying, inspiring and supporting new modes of business and innovation
- Providing a leading voice and coordinating framework for collective innovation in data-enabled science for the life sciences community
- Promoting sustainable and shared access to data, knowledge, tools, and services
- Facilitating information exchange through workshop development, community involvement and publications
- Identifying and cultivating opportunities to create essential solutions by implementing interdisciplinary and transdisciplinary approaches
- Coordinating and monitoring individual initiatives focused on translational impact

[back to top](#)

EarthCube

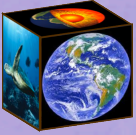
Transforming the conduct of research in the
geosciences

Presented by
Clifford Jacobs for the EarthCube Team
at

From GPS and Virtual Globes to Spatial Computing-2020



Sept 10-11, 2012



Vision=EarthCube Goal and Outcomes

Goal

To transform the conduct of research in geosciences by supporting the development of community-guided cyberinfrastructure to integrate data and information across the Geosciences.

Outcomes

Transform practices within the geosciences community spanning over the next decade

Provide unprecedented new capabilities to researchers and educators

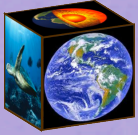
Vastly improve the productivity of community

Accelerate research on the Earth system

Provide a knowledge environment framework for the geosciences

Community Response

June 2011 to now



Social Network Site

<http://earthcube.ning.com/>

>1180 members to
the EarthCube
website

113 white paper
submission; 185
respondents to user
survey

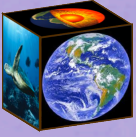
~70 expression of
interest emails

17 Community Groups

Unknown number of
hours of pro bono
contributions by the
community

Unprecedented view
of the pulse of the
geosciences
community

Significant International Engagement



NSF's Seven Modes of Success



We are proactive



We began with the end goal in mind



We prioritized EarthCube tasks



We emphasized a non-competitive & broadly inclusive process



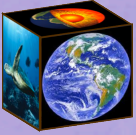
We listen to the community and shared our vision for EarthCube



We facilitated synergy within and across communities



We engaged and energized NSF colleagues in the process



Questions

What is the best way to take advantage of the current state of EarthCube?

Seven purple rectangular boxes for answers.

What portfolio of activities are the needed to engage the scientific community in the development/use of EarthCube?

Seven purple rectangular boxes for answers.

What are effectively mechanisms to engage our sister agencies and international partners in the EarthCube dialog?

Seven purple rectangular boxes for answers.

What are effective ways NSF can facilitate collaborative dialogs among the geosciences?

Seven blue rectangular boxes for answers.

What does it mean to change the community culture?

Seven light blue rectangular boxes for answers.